View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by FUOYE Journal of Engineering and Technology (FUOYEJET - Federal University...

A Comparative Analysis of Decision Tree and Bayesian Model for Network Intrusion Detection System

*[1]Bosede A. Ayogu, [2]Adebayo O. Adetunmbi and [3]Ikechukwu I. Ayogu

[1]Department of Computer Science, Federal University, Oye Ekiti, Ekiti State, Nigeria
[2]Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria
[3]Department of Computer Science, Federal Polytechnic, Idah, Kogi State, Nigeria
bosede.ayogu@fuoye.edu.ng | aoadetunmbi@futa.edu.ng | ig.ayogu@gmail.com

Abstract— Denial of Service Attacks (DoS) is a major threat to computer networks. This paper presents two approaches (Decision tree and Bayesian network) to the building of classifiers for DoS attack. Important attributes selection increases the classification accuracy of intrusion detection systems; as decision tree which has the advantage of generating explainable rules was used for the selection of relevant attributes in this research. A C4.5 decision tree dimensional reduction algorithm was used in reducing the 41 attributes of the KDD´99 dataset to 29. Thereafter, a rule based classification system (decision tree) was built as well as Bayesian network classification system for denial of service attack (DoS) based on the selected attributes. The classifiers were evaluated and compared using performance on the test dataset. Experimental results show that Decision Tree is robust and gives the highest percentage of successful classification than Bayesian Network which was found to be sensitive to the discretization techniques. It has been successfully tested that significant attribute selection is important in designing a real world intrusion detection system (IDS).

Keywords— Intrusion Detection System, Machine Learning, Decision Tree, and Bayesian Network

1 Introduction

Simultaneous accessibility of computer network by different individual all round the clock has made the systems and networks more vulnerable to different attacks, which is one of the critical issues in information technology. Attempts made to breach information security systems are on the rise with availability of tools that help accessing the vulnerability of network systems widely on the internet for free, thus making information and network security important aspect of today's information technology. Network traffic analysis is another aspect of computer security which grows more challenging each day; network traffic grows in volume and complexity in a manner that is both combinatorial and explosive due to cutting-edge technologies (Fatogun, 2012). Denial of Service Attack (DoS) is the most common attack affecting users on the network by simply denying legitimate users access to a machine.

Several network security measures like Firewall, Antivirus and Intrusion Detection Systems have been explored in attempts to mitigate network security breaches. The role of Intrusion Detection Systems (IDSs) in detecting anomalies and attacks in the network is becoming more important as firewall and antivirus have their own flaws. IDSs have been widely applied to overcome security threats in computer networks, it is a device or software application designed to check for malicious activity targeting a computer system or network (Mohit et al., 2017). It is not meant to replace prevention-based techniques such as authentication and access control, but to complement existing security measures and detect actions that bypass the security monitoring and control component of the system (Thomas et al., 2006). The two common approaches to intrusion detection are misuse based (signature based) and anomaly based approach (Nachiket, Durgesh, & Rajeshwar, 2018).

*Corresponding Author

Most of the IDS today are misuse based because of its simplicity and ability to detect known attacks. It implements the processes that match specific patterns in the observed traffic to established attack signatures. If a matching signature is found, the attack is detected. It has a very fast detection rate with less False Positive Rate (FPR), but can also lead to a high rate of False Negative as it requires regular updates for it to detect new attack as a new signature has to be developed whenever a new attack is discovered (Abhishek & Virender, 2017).

Anomaly detection is also one of the most frequently suggested approaches to detect new attacks without prior experience. It checks for strict deviations from normal profile of the network traffic and reports it as attack. The problem with anomaly detection mostly is the high level of false alarm rate due to the fact that the entire scope of the behaviour of an information system may not be covered during the learning phase (Saman, Najla,&Omar, 2010).

These shortcomings can be minimized by using informed scientific approaches for the selection of relevant attributes from the entire training dataset before they are been subjected to training in other to guarantee improved classification accuracy of the detection system (Revathi & Malathi, 2014). This paper employs a Decision tree feature selection approach and Bayesian classification model for building classifiers for DoS attack of the KDD'99 intrusion detection dataset.

2 Literature Review

Feature selection leads to simplification of the problem, faster, more accurate detection rates and increases the effectiveness of machine learning algorithms (Nachiket, Durgesh, & Rajeshwar, 2018; Adetunmbi, Adeola, & Daramola, 2011). Several researchers had applied these machine learning approaches in solving classification problems in intrusion detection and other areas. In 2009, Bayat et al applied a Decision Tree algorithms and Bayesian Network to the prediction of access to renal

transplant waiting list in a French healthcare network for optimizing the healthcare process. Their report showed that both models were of good sensitivity and specificity.

Farid et al. (2009) proposed a new learning approach using Bayesian classification and ID3 algorithm to address the problem of effective attributes selection and classification for network intrusion detection. The model exhibited high detection rate with reduced false alarm rate. Swamy & Lakshmi, (2012) applied decision tree technique to the construction of an intrusion detection system. C4.5, a descendant of ID3 and Improved decision tree was employed in the analysis of KDD'99 data set. The models were compared on the basis of false positive and true negative rates. The results showed that the improved Decision tree classified attacks better than the existing one.

Sharma, Jindal, & Kumar, (2012) applied a Bayesian network to build IDS which combined K2 algorithm, and Junction tree inference. The K2 algorithm was used to perform the state space for learning process, Bayesian for probabilistic decisions and junction tree for the analogy to generate the efficient tree for intrusion detection. Kruegel et al. (2003) proposed an event classification scheme based on Bayesian networks to mitigate the shortcomings of other existing models because it is seen to improve the aggregation of different model outputs and allow one to seamlessly incorporate additional information, as the existing models generate large number of false alarms because of the simplistic aggregation of model outputs in the decision phase and the lack of integration of additional information into the decision process.

Hema & Shyni (2015) proposed a Bayesian classification model for Denial of Service attacks to improve the detection rate and reduce the occurrence of false positive alarms in the system. An approach based on two Bernoulli and Multinomial Bayesian models which efficiently detects packet dropping attacks in Mobile Ad hoc Networks was described in Rmaythi et al. (2014). The research aimed at discovering secure paths between source and a destination by avoiding DoS attacks.

Sheikhan & Maryam (2010) employed Fuzzy association rule mining for the reduction of the features space from 41 to 31 features. The 31-attribute subsets were then used to learn a neural network employed in the determination of the attack category. Although, their system performed better in terms of detection rate, false alarm rate, and cost per example as compared to other machine learning methods, the input vectors could still be searched for lesser, more effective attribute set. Adetunmbi, Adeola, & Daramola, (2011) applied information gain, rough set degree of dependency and dependency ratio in selecting 20 most relevant features for DoS attacks from 41 features based on 10% Training set.

In 2009, Alowolodu applied a Genetic algorithm to differentiate between normal connection and attacks. The hypothesis search was restricted to nine fields and the features were selected based on its presumed importance.

Decision Tree is an effective feature selection machine learning technique use for detecting network intrusion with ability to analyze data and extracting the relevant features from large volume of data by identifying significant characteristics in the network that indicate malicious activities. The Intrusion detection based on decision tree is a straightforward classification method (Swamy & Lakshmi, 2012) as its simplicity and more direct interpretation tool for physicians has also made it to be more popular in the medical field (Bayat et al., 2009).

3 METHODOLOGY

3.1 DATA DESCRIPTION AND PRE-PROCESSING

For the purpose of this research, we have chosen KDD'99 intrusion detection dataset which was used in knowledge Discovery and Data mining (KDD) 1999 Cup competition. The dataset consists of input dataflow containing the details of the network connections, such as protocol type, connection duration, login type etc. Each data sample represents attribute or feature value of a class in the network data flow, and each class is labeled either as normal or as an attack with exactly one specific attack type. In total, 41 features for each connection record (discrete or continuous) plus one class label or decision attribute.

In our experiment, a total of 200,001 Patterns of DoS attack were extracted from the data set, covering 9 types of DoS attacks, out of which 89,409 was used for training and 3098 for testing instances respectively making a total of 92507. The categories involved are Normal, Smurf, Neptune, Mailbomb, Back, Teardrop, Processtable, Apache2, Land, and Pod. Table 1 shows the number of training and testing data for each category of attacks used for Decision tree and Bayesian Network.

Table 1. The Distribution of Instances in the Dataset

Categories	Occurrences in training	Occurrences in testing
Normal (NO)	500	501
Smurf (SM)	43614	508
Neptune (NE)	38814	601
Mailbomb (MB)	5000	500
Back (BC)	99	235
Teardrop (TD)	6	14
Processtable (PT)	506	338
Apache2 (AP2)	794	331
Land (LD)	8	10
Pod (PD)	68	60
TOTAL	89409	3098

3.2 EXPERIMENTAL PROCEDURE

The experiment is divided into two phases: training and testing using Decision Tree Algorithm and Bayesian Network Model. In the experiment, 41 attributes of the dataset were labeled in order as $A_1, A_2, \ldots\ldots A_{41}$. To select

the relevant attributes from training set, we have applied a C4.5 decision tree dimensionality reduction algorithm to reduce the 41attributes of the KDD'99 dataset to 29, which is the first phase of this experiment. The attributes considered most important were used to build a rule base classification system for detecting DoS attack.

Bayesian classifier belongs to the group of Machine Learning that learns using discrete attributes, and as such, equal-bin discretization technique was used on the continuous attributes of the reduced dataset before being subjected to training. The distribution for training and testing data for both models is shown in Table 1.

In the second phase of the experiment, the comparison of the results is done on the basis of accuracy and error rate to classify the categories as normal, smurf, Neptune, mailbomb, Back, Land, Teardrop, processtable, pod, and apache2. All experiments were performed on windows 7 running Intel core 2 Duo processor, 2.0GHz of CPU and 2GB of RAM. C-Sharp and DotNet framework were used in developing the system.

3.3 Learning Algorithms

3.3.1 Decision Tree Learning Algorithm

Decision tree is composed of three basic elements: a decision node which specifies a test attribute (Uttam & Satyendra, 2013). An edge or a branch corresponds to the one of the possible attribute values or outcomes. A leaf which is also named an answer node contains the class to which the object belongs. To reduce the dimension of input features, a decision tree applies a divide and conquers method by partitioning tuples or records that have mixture of classes to subsets, so that each subset would belong to a single class. The decision tree procedure as described in a number of sources, including (Farid, Harbi, & Rahman, 2010; Vaishali et al., 2014; Rai, Devi, & Guleria, 2016).

Given a training set T of m-tuples which contains n-features/attributes $F_i = \{ f_1, f_2, \ldots . f_n \}$ with feature values $X_i = \{ x_1, x_2, \ldots . x_n \}$, which may be discrete or continuous. Each tuple in the training set is associated with a particular class and class labels $C_i = \{ c_1, c_2, \ldots . c_k \}$; the divide and conquer algorithm calculates the information gain for each feature from the training set T, selecting one tuple at random from a training set of n-tuples and announcing that it belongs to some class c_i. The procedure for partitioning begins as described below.

Step 1. Start
Step 2. Get a training set T of m-tuples with n-features/attributes associated with particular classes
Step 3. Place the best attribute of the dataset at the root of the tree
Step 4. Split the training set into subsets in such a way that each subset contains data with the same value for attribute using information gain.
Step 5. For a continuous attribute, create a threshold and then split the list into those whose attribute value is above the threshold and those that are less or equal to it.
Step 6. Starting from the root node select attribute with the maximum information gain which is the next node in the tree
Step 7. Repeat steps3 and 4 on each subset until leaf nodes are found in all the branches of the tree
Step 8. Generate a distinct rule for classifying each attack types using gain ratio criterion.
Step 9. Stop
Step 10. Return root

3.3.2 Bayesian Network Learning Algorithm

Bayesian Network is also one of the machine learning techniques for detecting intrusion. It is a probabilistic approach to detect the intrusion; it allows the incorporation of additional information to improve the aggregation of different models (Sharma, Jindal, & Kumar, 2012), but its performance depends on the discretization technique being used.

Given a tuple, X_i, such that X_i is continuous value, then there is need for normalization as Bayesian classifier learns with discrete attributes. In order to find the interval for each attribute, the maximum value is subtracted from the minimum value divided by the number of partition using equal bin discretization technique (Chaouki & Saoussen, 2017).

Thereafter, the Bayesian classifier will predict that X_i belongs to the class having the highest posterior probability, conditioned on X_i. That is, the naïve Bayesian classifier predicts that tuple X_i belongs to the class C_i if and only if $P(C_i|X_i) > P(C_j|X_i)$ for $1 \leq j \leq k, j \neq i$. Thus, we maximize conditional probability $P(C_i|X_i)$over class C_i. The class C_i for which $P(C_i|X_i))$ is maximized is called the maximum posterior hypothesis (Petre Ray 2015) which is calculated from $P(C_i)$, $P(X)$and$P(X|C_i)$. Below is an algorithm for Bayesian network.

Input: The reduced dataset
Output: Probabilities of attack and normal records stored in a file
Step 1. If feature X contains continuous value X_i, then normalize X_i
Step 2. Read processed dataset file for each cluster head
Step 3. Call naive Bayesian classifier program for training the classifier for DoS attack detection, and store the information into a file
Step 3. Test this file with the classifier model and write them to an output file.

4 Result Discussion

This section describes the experimental results and performance of the models on the basis of accuracy and error rate. The attributes were successfully reduced from 41 to 29, and the categories were classified as normal, smurf, Neptune, mailbomb, Back, Land, Teardrop, processtable, pod, and apache2 under the 29 reduced attributes. All the 501 normal records that were tested against the DTM were correctly classified while 494 records were correctly classified by BNM. DTM also classified all the 508 smurf attack correctly while BNM predicted 504 correctly out of the 508, etc. Table 2 shows the comparison performance of the models for each category in percentage. Altogether 3098 records were tested out of which 3073 were correctly predicted by Decision Tree which has 99.19 % Accuracy and 0.8%

Error Rate, whilst 3036 instances were correctly predicted by Bayesian Network Model with 97.99%. Accuracy and 2% error rate as shown in Table 3, while figure 1 shows the graphical representation of the comparison.

5 Conclusion and Future Work

Feature subset construction is an important exercise in the development of efficient IDS. This paper devised a 29-attriubute /feature subset from the 41-attributes of KDD'99 intrusion detection dataset using C4.5 Decision tree model. A decision tree rule based classification system was built based on the attributes selected. The result shows that both models performed well in terms of the metrics used in the experiment but decision tree still has a better performance than the Bayesian network.

Table 2. Detection Rate per Category

Category	DTM	BNM
SM	100	99.21
PD	96.66	91.66
BC	100	95.32
AP2	99.70	95.77
NE	100	97.67
LD	100	100
PT	94.67	98.52
MB	100	99.60
TP	100	100
NO	99.20	98.60

Table 3. A Summary of Models Performance

	DTM	BNM
Accuracy (%)	99.19	97.99
Error Rate (%)	0.8	2.0

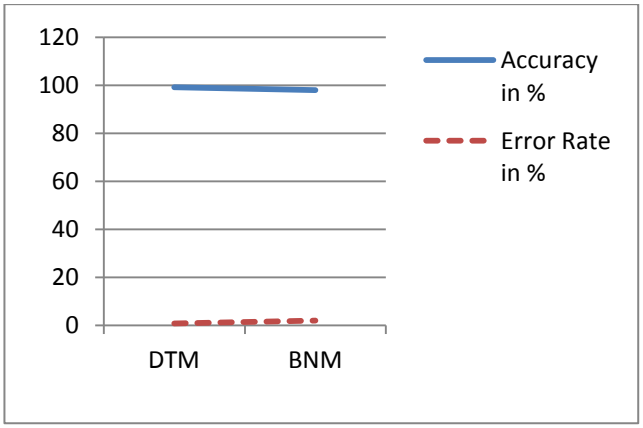

Fig. 1: Comparison of Accuracy and Error rate for DTM and BNM

This research improves detection rate of DoS attack on computer network as well as minimizes the computational complexity of the classifiers. Further work will focus on investigating other data reduction methods.

Acknowledgment

My profound gratitude goes to God Almighty for seeing me through the course of this research and also to my co-authors for their contributions towards the successful completion of this work.

References

Abhishek, V., & Virender, R. (2017). Statistical Analysis of CIDDS-001 Dataset for Network Intrusion Detection Systems Using Distanced-Based Machine Learning. 6th International Conference on Smart Computing and Communications, ICSCC 2017.

Adetunmbi A. O., Adeola O.S, & Daramola, O. A. (2011). Relevance Features Selection for Intrusion Detection. Intelligent Automation and System Engineering, Lecture Notes in Electrical Engineering, (*Boston Springer*), vol. 103 pp. 407 – 418.

Alowolodu, O. D. (2009). Intrusion Detection System Using Genetics Algorithm to differentiate between normal and attacks. A master thesis in the Department of Computer Science, Federal University of Technology Akure, Nigeria.

Bayat, S., Cuggia, M., Rossille, D., Kessler, M., & Frimat, L. (2009). Comparison of Bayesian Network and Decision Tree Methods for Predicting Access to the Renal Transplant Waiting List. Medical Informatics in a United and Healthy –Europe.

Chaouki, K., & Saoussen, K. (2017). A GA-LR wrapper approach for feature selection in network intrusion detection. Retrieved from www.sciencedirect.com.

Farid, D. M., Darmont, J., Harbi, N., Haa, N. H., & Rahman, M. Z. (2009). Adaptive Network Intrusion Detection Learning: Attributes Selection and Classification. In: World Academy of Science, Engineering and Technology 60.

Farid, D. M., Harbi, N., & Rahman, M. Z. (2010). Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection; International Journal of Network Security & its Applications. Pp 12-25.

Fatogun, B.A. (2012). Denial of Service Attack Detection Using Machine Learning Techniques. Unpublished master thesis in the Department of Computer Science, Federal University of Technology Akure, Nigeria.

Hema, V., & Shyni, E. C. (2015). DoS Attack Detection Based on Naive Bayes Classifier. *Middle-East Journal of Scientific Research (Sensing, Signal Processing and Security)*, Pp. 398-405.

KDDCup 1999 Data: http://kdd.ics.uci.edu/databases/kddcup99.

Kruegel,C., Mutz, D., Robertson, W., & Valeur, F. (2003). Bayesian Event Classification for Intrusion Detection. Conference proceeding in Computer Security Applications. Pp 14-23. IEEE.

Mohit, T., Raj, K., Akash, B., & Jai, K. (2017). Intrusion Detection System. *International Journal of Technology Research and Applications, Issue 2, Vol., 5 Pp 38-44.*

Nachiket, S., Durgesh, S., & Rajeshwar, S. (2018). Feature Classification and Outlier Detection to Increased Accuracy in Intrusion Detection System. *International Journal of Applied Engineering Research Vol. 13, No. 10.*

Petre, R. (2015). Enhancing Forecasting Performance of Naive Bayes Classifier with Discretization Techniques. *Database System Journal*. Vol. 4, No. 2.

Rai, K., Devi, M. S., & Guleria, A. (2016). Decision Tree Based Algorithm for Intrusion Detection: *International Journal of Advanced Networking and Applications. Vol. 7 Iss 4. Pp: 2828-2834.*

Revathi, S., & Malathi, A. (2014). Network Intrusion Detection Using Hybrid Simplified Swarm Optimization and Random Forest Algorithm on Nsl-Kdd Dataset: *International Journal of Engineering and Computer Science Vol. 3, No 3873-3876.*

Rmaythi, M., Begriche, Y., Khatoun, R., Khoukhi, L., & Gaiti, D. (2014). Denial of Service(DoS) Attacks Detection in MANETs using Bayesian Classifiers. In 2014 IEEE 21st Symposium on Communications and Vehicular Technology in the Benelux (SCVT), pp. 7-12).

Saman, M. A., Najla, B. A., & Omar, Z. (2010). Identify Features and Parameters to Devise an Accurate Intrusion Detection System Using Artificial Neural Network. *International Journal of Computer and Information Engineering. Vol: 4, No: 10.*

Sharma, M., Jindal, K., & Kumar, A. (2012). Intrusion Detection System Using Bayesian Approach for Wireless Network. *International Journal of Computer Applications. Vol. 48, No 5.*

Sheikhan, M., & Maryam, S. (2010). Detection Based on Feature Selection by Fuzzy Association Rule Mining. In: World Applied Sciences Journal 10 (Special Issue of Computer & Electrical Engineering): 32-40, 2010. ISSN 1818-4952. www.idosi.org/wasj/wasj10(C&EE)2010.htm

Swamy, K. V. R., & Lakshmi, K. S. (2012). Network Intrusion Detection Using Improved Decision Tree Algorithm. *International Journal of Computer Science and Information Technologies. Vol., 3. Pp 4971-4975.*

Thomas, M. C., Geng-Sheng, K., Zheng-Ping, L., & Guo-Mei, Z. (2006). Intrusion Detection in Wireless Mesh Networks. http://lyle.smu.edu/~tchen/papers/intrusion-detection-mesh.pdf.

Uttam, B. J., & Satyendra, V. (2013). Decision Tree Based Intrusion Detection System Using Wrapper Approach. *International Journal of Advances In Engineering & Technology. Vol. 6, Iss 5, Pp.2187-2195.*

Vaishali, K., & Sangita, S. C. (2014). Improved Intrusion Detection System Using C4.5 Decision Tree and Support Vector Machine. *International Journal of Computer Science and Information Technologies, Vol. 5, Issue 2, Pp. 1463-1467.*